(19) **RU** (11) **2 166 207** (13) **C2**

(51) Int. Cl.[7] **G 06 F 17/20, 17/21, 17/22, 17/27, 17/28**

RUSSIAN AGENCY
FOR PATENTS AND TRADEMARKS

(12) ABSTRACT OF INVENTION

(21), (22) Application: **99106483/09, 08.04.1999**

(24) Effective date for property rights: **08.04.1999**

(46) Date of publication: **27.04.2001**

(98) Mail address:
115573, Moskva, Orekhovyj bul'var, d.39, korp.1, kv.145, Linniku L.N.

(71) Applicant:
Zakrytoe aktsionernoe obshchestvo "Abi Programmnoe obespechenie"

(72) Inventor: **Anisimovich K.V.**, Tereshchenko V.V., Jan D.E.

(73) Proprietor:
Zakrytoe aktsionernoe obshchestvo "Abi Programmnoe obespechenie"

(54) METHOD FOR USING AUXILIARY DATA ARRAYS IN CONVERSION AND/OR VERIFICATION OF CHARACTER-EXPRESSED COMPUTER CODES AND RESPECTIVE SUBPICTURES

(57) Abstract:

FIELD: electronics. SUBSTANCE: method involves retrieval of significant units of subpictures to be recognized that incorporate n component pixels, where n is chosen within $1 \le n \le 10^3$ range. Sets of subpictures to be verified that have n_1 pixels are selected, where n_1 is chosen within $1 \le (n_1+n)/n \le 2$ range. Significant units that differ from selected sets of subpictures are retrieved in auxiliary data array with error ε chosen within $0 \le \varepsilon \le (\alpha n_1 - 1)/n_1$ range, where α is experimental factor within $0,6 \le \alpha \le 1,2$ range selected as function of rate of occurrence of any i-th significant unit in permissible set of significant units which is defined as repetitive quantity n_2 of particular significant units related to total quantity n_3 of significant units in their permissible set. Pixels that do not coincide with characters equivalent to them in location in significant units found in the course of retrieval are detected in recognized significant units equivalent to them in location and replaced by characters of respective location retrieved from significant units found. Additional array of dynamic raster standards of computer codes is formed as part of recognizable significant units and auxiliary data array is converted, bearing in mind preceding operations, until total error ε_3 of method chosen relative to intermediate error ε_1 is reduced within $1 \le (\varepsilon_1+\varepsilon_3)/\varepsilon_1 \le 2$. range. EFFECT: reduced conversion and/or verification error.

RU 2166207 C2

RU 2166207 C2

(19) **RU** (11) **2 166 207** (13) **C2**

(51) МПК[7] **G 06 F 17/20, 17/21, 17/22, 17/27, 17/28**

РОССИЙСКОЕ АГЕНТСТВО
ПО ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ РОССИЙСКОЙ ФЕДЕРАЦИИ

(21), (22) Заявка: **99106483/09, 08.04.1999**

(24) Дата начала действия патента: **08.04.1999**

(46) Дата публикации: **27.04.2001**

(56) Ссылки: **WO 98/24036 A1, 04.06.1998. RU 2112273 C1, 27.05.1998. WO 96/34348 A1, 31.10.1996. WO 98/00794 A1, 08.01.1998. RU 2096825 C1, 20.11.1997. US 5477451 A1, 19.12.1995.**

(98) Адрес для переписки:
115573, Москва, Ореховый бульвар, д.39, корп.1, кв.145, Линнику Л.Н.

(71) Заявитель:
Закрытое акционерное общество "Аби Программное обеспечение"

(72) Изобретатель: **Анисимович К.В., Терещенко В.В., Ян Д.Е.**

(73) Патентообладатель:
Закрытое акционерное общество "Аби Программное обеспечение"

(54) СПОСОБ ИСПОЛЬЗОВАНИЯ ВСПОМОГАТЕЛЬНЫХ МАССИВОВ ДАННЫХ В ПРОЦЕССЕ ПРЕОБРАЗОВАНИЯ И/ИЛИ ВЕРИФИКАЦИИ КОМПЬЮТЕРНЫХ КОДОВ, ВЫПОЛНЕННЫХ В ВИДЕ СИМВОЛОВ, И СООТВЕТСТВУЮЩИХ ИМ ФРАГМЕНТОВ ИЗОБРАЖЕНИЯ

(57)

Изобретение относится к области электроники и предназначено, например, для использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов, выполненных в виде символов, и соответствующих им фрагментов изображения. Техническим результатом является снижение погрешности преобразования и/или верификации. Способ заключается в том, что производят выработку смысловых единиц распознаваемых фрагментов изображения, содержащих n составляющих их элементов, где n выбирают в пределах $1 \le n \le 10^3$. В отобранных выборках выделяют подлежащие верификации совокупности их фрагментов изображения, содержащие n_1 элементов, где n_1 выбирают в пределах $1 \le (n_1+n)/n \le 2$. Осуществляют поиск во вспомогательном массиве данных смысловых единиц, отличающихся от выделенных совокупностей фрагментов изображения, с погрешностью ε, выбираемой в пределах $0 \le \varepsilon \le (\alpha n_1 - 1)/n_1$, где α - экспериментальный коэффициент в пределах $0{,}6 \le \alpha \le 1{,}2$, выбираемый в зависимости от частоты f_i появления любой смысловой i-й единицы в допустимом множестве смысловых единиц, которую определяют как количество n_2 повторений конкретной смысловой единицы, соотнесенное с общим количеством n_3 смысловых единиц в допустимом множестве смысловых единиц. Выявляют в распознанных смысловых единицах элементы, которые не совпадают с эквивалентными им по месту расположения символами в смысловых единицах, найденных в процессе поиска, и производят их замену соответствующими им по месту расположения символами из найденных смысловых единиц. Формируют дополнительный массив динамических растровых эталонов компьютерных кодов элементов в составе распознаваемых смысловых единиц и с учетом предшествующих операций преобразуют вспомогательный массив данных до уменьшения итоговой погрешности ε_3 способа, которую выбирают по отношению к промежуточной погрешности ε_1 в пределах $1 \le (\varepsilon_1+\varepsilon_3)/\varepsilon_1 \le 2$.

RU 2166207 C2

RU 2166207 C2

Изобретение относится к области электроники и может быть применено, например, для использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов, выполненных в виде символов, и соответствующих им фрагментов изображения.

Известен способ использования вспомогательных данных в процессе преобразования компьютерных кодов и соответствующих им фрагментов изображения, включающий производимое человеком и/или заменяющим его устройством, и/или компьютерной программой использование вспомогательных данных, привлекаемых для распознавания соответствующих им оригиналов [Patent USA N 5153927: Character reading system and method., МПК Oct. 6, 1992.].

Известен также способ использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов и соответствующих им оригиналов, заключающийся в осуществляемом компьютерной программой использовании вспомогательных массивов данных, привлекаемых для распознавания соответствующих им оригиналов [Руководство пользователя Fine Reader 4.0 © ABBYY Software House, M. , 1998. Казанский производственный комбинат программных средств. Заказ Ф-377].

Недостатком известных способов являются относительно низкие их функциональные и технические характеристики, в том числе высокие значения достигаемых погрешностей преобразования.

Решаемой изобретением задачей является совершенствование способов использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов, выполненных в виде символов, и соответствующих им фрагментов изображения с достижением технического результата в виде снижения погрешности преобразования и/или верификации.

Для удобства и однозначного понимания целесообразно привести расшифровки и определения используемых далее обозначений, символов и/или терминов.

Оригинал - преобразуемая информация, материализованная преимущественно в виде совокупности компьютерных кодов, соответствующих исходному объекту, например распознаваемому фрагменту изображения.

Компьютерный код (например, символ) - компьютерное представление некоторого фрагмента информаций (в частности, символьной).

Процесс распознавания - процесс обработки системой распознавания введенного в компьютер графического изображения некоторого символа, в результате чего система распознавания приписывает изображению компьютерный код этого символа.

Процесс верификации - производимое человеком и/или заменяющим его устройством, и/или компьютерной программой сличение (определение адекватности) компьютерных кодов (символов) с графическим изображением, введенным в компьютер.

Допустимое множество смысловых единиц включает в себя всю совокупность вероятных для распознавания наборов смысловых единиц.

Смысловая единица - это совокупность компьютерных кодов, соответствующих ориентированному на какое-либо практическое использование образу, например букве, слову, символу, графическому элементу, логической операции, их совокупности и др.

Вспомогательный массив данных - это произвольным образом сформированная совокупность электронных кодов смысловых единиц, охватываемых, в частности, допустимым множеством смысловых единиц.

Погрешность соответствия ε между исходными смысловыми единицами и соответствующими им смысловыми единицами объема n_1 в дополнительном массиве данных, определяется как допустимое число $\triangle n_1$ несовпадающих в них элементов, соотнесенное с n_1: $\varepsilon = \triangle n_1/n_1$.

Частота f_i появления любой смысловой i-й единицы в допустимом множестве смысловых единиц определяется как количество n_2 повторений конкретной смысловой единицы, соотнесенное с общим количеством из смысловых единиц в допустимом множестве смысловых единиц: $f_1 = n_2^1/n_3$.

Погрешность ε_1 вспомогательного массива данных по отношению к допустимому множеству смысловых единиц определяется, как вероятность не нахождения в массиве данных элемента n_j, соотнесенного с общим количеством смысловых единиц n_4 во вспомогательном массиве данных.

Погрешность ε_2 преобразования определяется как количество n_5 ошибочно преобразованных элементов, соотнесенные с общим количеством n_6 элементов в преобразуемом наборе смысловых элементов из их допустимого множества: $\varepsilon_2 = n_5/n_6$.

Погрешность ε_3 определяется как итоговая погрешность преобразования.

Дополнительный массив динамических растровых эталонов - это совокупность элементов смысловых единиц, формируемая в процессе преобразования для уменьшения погрешностей ε_1, ε_2.

В качестве кратких сведений, раскрывающих сущность изобретения следует отметить, что достигаемый технический результат обеспечивают с помощью предложенного способа использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов, выполненных в виде символов, и соответствующих им фрагментов изображения, заключающегося в том, что производят выборку смысловых единиц распознаваемых фрагментов изображения, содержащих n_1 составляющих их элементов, где n_1 - выбирают в пределах $1 \le n \le 10^3$. В отобранных выборках выделяют подлежащие верификации совокупности их фрагментов изображения, содержащие n_1 элементов, где n_1 выбирают в пределах $1 \le (n_1+n)/n \le 2$. Осуществляют поиск во вспомогательном массиве данных смысловых единиц, отличающихся от выделенных совокупностей фрагментов изображения, с

RU 2166207 C2

RU 2166207 C2

погрешностью ε выбираемой в пределах $0 \le \varepsilon \le (\alpha n_1 - 1)/n_1$. Здесь α - экспериментальный коэффициент в пределах $0,6 \le \alpha \le 1,2$, выбираемый в зависимости от частоты f_i появления любой смысловой i-й единицы в допустимом множестве смысловых единиц, которую определяют как количество n_2 повторений конкретной смысловой единицы, соотнесенное с общим количеством n_3 смысловых единиц в допустимом множестве смысловых единиц.

Выявляют в распознанных смысловых единицах элементы, которые не совпадают с эквивалентными им по месту расположения символами в смысловых единицах, найденных в процессе поиска, и производят их замену соответствующими им по месту расположения символами из найденных смысловых единиц. Формируют дополнительный массив динамических растровых эталонов компьютерных кодов элементов в составе распознаваемых смысловых единиц количеством n_7, величину которого выбирают в пределах $1 \le (n_2 + n_5 + n_6 + \beta n_7 + n_3)/n_3 \le 6,3$. Здесь β - - экспериментальный коэффициент в пределах $0,4 \le \beta \le 1,3$, выбираемый в зависимости от задаваемых погрешности ε_1 вспомогательного массива данных по отношению к допустимому множеству смысловых единиц, определяемой как вероятность не нахождения в массиве данных элемента n_j, в общем количестве смысловых единиц n_4 во вспомогательном массиве данных, и погрешности ε_2 преобразования, определяемой как количество n_5 ошибочно преобразованных элементов, соотнесенное с общим количеством n_6 элементов в преобразуемом наборе смысловых элементов из их допустимого множества. Затем с учетом предшествующих операций преобразуют вспомогательный массив данных до уменьшения итоговой погрешности ε_3 способа, которую выбирают по отношению к погрешности ε_1 в пределах $1 \le (\varepsilon_1 + \varepsilon_3)/\varepsilon_1 \le 2$.

При изложении сведений, подтверждающих возможность осуществления изобретения целесообразно более подробно описать предложенный способ использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов и соответствующих им фрагментов изображения. Детально целесообразно остановиться только на существенных особенностях осуществления операций предложенного способа, заключающегося в том, что производят выборку смысловых единиц распознаваемых фрагментов изображения, содержащих n составляющих их элементов, где n - выбирают в пределах $1 \le n \le 10^3$. Смысловыми единицами могут быть в произвольном случае буквы, математические и другие символы, отдельные слова, предложение, графические элементы, а также их любые сочетания. В отобранных выборках выделяют подлежащие верификации совокупности их фрагментов изображения, содержащие n_1 элементов, где n_1 выбирают в пределах $1 \le (n_1+n)/n \le 2$. Осуществляют поиск во вспомогательном массиве данных смысловых единиц, отличающихся от выделенных совокупностей фрагментов изображения, с погрешностью ε, выбираемой в пределах $0 \le \varepsilon \le (\alpha n_1 - 1)/n_1$. Здесь α - экспериментальный коэффициент в пределах $0,6 \le \alpha \le 1,2$, выбираемый в зависимости от частоты f_i появления любой смысловой i-й единицы в допустимом множестве смысловых единиц, которую определяют как количество n_2 повторений конкретной смысловой единицы, соотнесенное с общим количеством n_3 смысловых единиц в допустимом множестве смысловых единиц. Фрагментами могут быть как смысловые единицы в целом, так и их части, ориентированные, например, на автономное применение. Погрешность преобразования в основном связана с качеством исходных графических изображений, которое определяется, в частности, тем, что предъявляют для распознавания, например, изготовленное на ксерокопировальном аппарате изображение, факсограмму, машинописный или рукописный текст.

Выявляют в распознанных смысловых единицах элементы, которые не совпадают с эквивалентными им по месту расположения символами в смысловых единицах, найденных в процессе поиска, и производят их замену соответствующими им по месту расположения символами из найденных смысловых единиц. Формируют дополнительный массив динамических растровых эталонов компьютерных кодов элементов в составе распознаваемых смысловых единиц количеством n_7, величину которого выбирают в пределах $1 \le (n_2 + n_5 + n_6 + \beta n_7 + n_3)/n_3 \le 6,3$. Здесь β - экспериментальный коэффициент в пределах $0,4 \le \beta \le 1,3$, выбираемый в зависимости от задаваемых погрешности ε_1 вспомогательного массива данных по отношению к допустимому множеству смысловых единиц, определяемой как вероятность не нахождения в массиве данных элемента n_j, в общем количестве смысловых единиц n_4 во вспомогательном массиве данных, и погрешности ε_2 преобразования, определяемой как количество n_5 ошибочно преобразованных элементов, соотнесенное с общим количеством n_6 элементов в преобразуемом наборе смысловых элементов из их допустимого множества.

Процесс построения динамических растровых эталонов целесообразно определить как производимое человеком и/или заменяющим его устройством, и/или компьютерной программой построение дополнительного массива данных, используемых для ускорения процесса распознавания. Динамический растровый эталон - это дополнительный массив данных, создаваемый и используемый для ускорения процесса распознавания. Понятие "динамический" отражает изменяемый характер создаваемых эталонов, то есть означает, что в процессе использования предложенного способа постоянно изменяют совокупность построенных эталонов пополнением ее новыми эталонами, видоизменением существующих и исключением ненужных, а понятие

RU 2166207 C2

RU 2166207 C2

"растровый" характеризует их особенности выполнения в виде совокупности элементов, образующих, например, периодическую структуру. Создать эталон - значит для каждого встречающегося в тексте символа записать в память растровой подсистемы пару: точечное изображение символа и его название (т.е. какую буквы это изображение представляет).

Затем с учетом предшествующих операций преобразуют вспомогательный массив данных до уменьшения итоговой погрешности ε_3 способа, которую выбирают по отношению к погрешности ε_1 в пределах $1 \leq (\varepsilon_1+\varepsilon_3)/\varepsilon_1 \leq 2$. На практике возможно использование и отдельных логически завершенных совокупностей операций предложенного способа. Если в результате выделения в соответствии с приведенными аналитическими соотношениями необходимых величин получают дробные, отрицательные значения и какие-либо другие значения, некорректные исходя из условий возможности их дальнейшего использования, то их исключают из рассмотрения и/или автоматически удаляют.

В качестве примера практического выполнения заявленного способа использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов и соответствующих им фрагментов изображения, целесообразно привести следующий, реализованный в последних версиях системы оптического распознавания текстов FineReader. В процессе реализации способа производят выборку смысловых единиц распознаваемых оригиналов, содержащих n составляющих их элементов, где n - выбирают в пределах $1 \leq n \leq 20$. В отобранных выборках выделяют подлежащие верификации совокупности их фрагментов, содержащие n_1 элементов, где n_1 выбирают из условия $1,8 \leq (n_1+n)/n \leq 2$. Осуществляют поиск во вспомогательном массиве данных смысловых единиц с погрешностью ε отличающихся от выделенных совокупностей фрагментов, выбираемой в пределах $\varepsilon \leq 0,1$ при $\alpha = 0,9$ f_i = 0,01 - 0,1. Выявляют в распознанных смысловых единицах элементы, которые не совпадают с эквивалентными им по месту расположения символами в смысловых единицах, найденных в процессе поиска, и производят их замену соответствующими им по месту расположения символами из найденных смысловых единиц. Формируют дополнительный массив динамических растровых эталонов компьютерных кодов элементов в составе распознаваемых смысловых единиц количеством n_7, величину которого по отношению к общему количеству n_3 смысловых единиц в допустимом множестве смысловых единиц выбирают из условия $n_7/n_3 = 0,9$ при β = 1,1, $\varepsilon_1 = 0,05$ и $\varepsilon_2 = 0,05$, пренебрегая в конкретном случае влиянием n_2, n_5, и n_6 на величину n_7. Преобразуют в результате вспомогательный массив данных до уменьшения погрешности ε_3 по отношению к погрешности ε_1 из условия $(\varepsilon_1+\varepsilon_3)/\varepsilon_1 = 1,2$.

Компьютерный код в заявленном объекте, как уже указывалось, - это преобразуемая компьютером совокупность электромагнитных сигналов, адекватно соответствующих распознаваемым исходным символам или любым другим распознаваемым фрагментам исходной информации. Каждый из эталонов совокупности динамических растровых эталонов, образующих периодическую структуру, представляет собой, например, упорядоченный набор электромагнитных сигналов или соответствующих рельефно намагниченных фрагментов жесткого диска. Динамические свойства растровых эталонов определяют временными параметрами их преобразования.

В отношении технических средств, необходимых для реализации заявленного способа, целесообразно в дополнении к вышеизложенному отметить, что ими могут быть как специализированные функциональные блоки, так и функциональные узлы компьютера, управляемые задаваемой системой команд. В частности, некоторые операции осуществляются математическим сопроцессором центрального процессора системного блока компьютера под управлением специализированных для операций с массивами данных и статистических вычислений функциональных программных блоков (библиотек команд, эталонов и других данных), производящих выборку и сортировку списков эталонов. Сами списки находятся либо в оперативной памяти (ОЗУ), либо на дисковом носителе компьютера и управляются системными библиотеками команд операционной среды. Под заменяющим человека устройством подразумевается любое устройство, которое может на необходимом для осуществления способа уровне выполнить операции, которые ранее выполнял или которые может выполнить человек. На практике техническими средствами реализации способа построения динамических растровых эталонов компьютерных кодов в процессе распознавания соответствующих им оригиналов могут являться, в частности, система состоящая из сканера, компьютера с загруженной в оперативную память программой сканирования, программой Fine Reader, подсистемой синхронизации компьютерных кодов, а также монитора, либо печатающего устройства и манипулятора для контроля и управления процессом.

Соответствие критерию промышленная применимость предложенного способа также доказывается отсутствием в заявленных притязаниях каких-либо практически трудно реализуемых признаков и известностью средств для их осуществления.

Указанные в формуле изобретения отличия, как уже отмечалось, дают основание сделать вывод о новизне предложенного технического решения, а совокупность испрашиваемых притязаний - о неочевидности их создания или об их изобретательном уровне, что доказывается также вышеприведенным описанием способа. Практическое использование способа обеспечивает достижение вышеуказанного технического результата взаимосвязанной совокупностью существенных признаков и особенностей, отраженных в формуле

RU 2166207 C2

RU 2166207 C2

изобретения. Особенности использования способа и других объектов, не отраженные в описании, общеизвестны и не являются предметом изобретения.

Кроме указанного выше технического результата, практическое осуществление заявленного объекта позволяет существенно расширить возможности его использования применительно, например, к различным документам, заполняемым рукописными символами, либо документам плохого качества.

Формула изобретения:

Способ использования вспомогательных массивов данных в процессе преобразования и/или верификации компьютерных кодов, выполненных в виде символов, и соответствующих им фрагментов изображения, заключающийся в том, что производят выборку смысловых единиц распознаваемых фрагментов изображения, содержащих n составляющих их элементов, где n выбирают в пределах $1 \le n \le 10^3$, в отобранных выборках выделяют подлежащие верификации совокупности их фрагментов изображения, содержащие n_1 элементов, где n_1 выбирают в пределах $1 \le (n_1 + n)/n \le 2$, осуществляют поиск во вспомогательном массиве данных смысловых единиц, отличающихся от выделенных совокупностей фрагментов изображения, с погрешностью ε, выдираемой в пределах $0 \le \varepsilon \le (\alpha n_1 - 1)/n_1$ где α - экспериментальный коэффициент в пределах $0,6 \le \alpha \le 1,2$, выбираемый в зависимости от части fi появления любой смысловой i-ой единицы в допустимом множестве смысловых единиц, которую определяют как количество n_2 повторений конкретной смысловой единицы, соотнесенное с общим количеством n_3 смысловых единиц в допустимом множестве смысловых единиц, выявляют в распознанных смысловых единицах элементы, которые не совпадают с эквивалентными им по месту расположения символами в смысловых единицах, найденных в процессе поиска, и производят их замену соответствующими им по месту расположения символами из найденных смысловых единиц, формируют дополнительный массив динамических растровых эталонов компьютерных кодов элементов в составе распознаваемых смысловых единиц количеством n_7, величину которого выбирают в пределах $1 \le (n_2 + n_5 + n_6 + \beta n_7 + n_3)/n_3 \le 6,3$, где β - экспериментальный коэффициент в пределах $0,4 \le \beta \le 1,3$, выбираемый в зависимости от задаваемых погрешности ε_1 вспомогательного массива данных по отношению к допустимому множеству смысловых единиц, определяемой как вероятность ненахождения в массиве данных элемента n_j, в общем количестве смысловых единиц n_4 во вспомогательном массиве данных, и погрешности ε_2 преобразования, определяемой как количество n_5 ошибочно преобразованных элементов, соотнесенное с общим количеством n_6 элементов в преобразуемом наборе смысловых элементов из их допустимого множества, и преобразуют вспомогательный массив данных до уменьшения погрешности ε_3 способа, которую выбирают по отношению к погрешности ε_1 в пределах $1 \le (\varepsilon_1 + \varepsilon_3)/\varepsilon_1 \le 2$.

RU 2166207 C2

RU 2166207 C2